

文化偏见在许多大语言模型（LLMs）中普遍存在，主要是由于缺乏代表不同文化的数据。通常，文化数据集和基准是通过提取现有数据集的子集或从维基百科和社交媒体等平台聚合数据来构建的。然而，这些方法高度依赖于真实世界的数据和人工标注，使其成本高昂且难以扩展。受到社会传播认知理论的启发，本文介绍了CulturePark，一个基于LLM的多智能体通信框架，用于文化数据的收集。CulturePark模拟跨文化的人类交流，通过基于LLM的智能体扮演不同文化中的角色，生成高质量的跨文化对话，encapsulating 人类的信仰、规范和习俗。通过使用CulturePark，我们生成了41,000个文化样本，用于微调八种文化特定的LLM。我们在三个下游任务上评估了这些模型：内容审核、文化对齐和文化教育。结果表明，在内容审核方面，我们基于GPT-3.5的模型在41个数据集上与GPT-4相当或表现更好。在文化对齐方面，我们的模型在Hofstede的VSM 13框架上超越了GPT-4。此外，在文化教育方面，我们的模型在学习效果和用户体验上均优于GPT-4。CulturePark是解决文化偏见和推动人工智能民主化的重要一步，突显了文化包容性数据在模型训练中的关键作用。代码已发布在<https://github.com/Scarelette/CulturePark>。

CulturePark: Boosting Cross-cultural Understanding in Large Language Models

Cheng Li*

Institute of Software, CAS
chenglicat0228@gmail.com

Damien Teney

Idiap Research Institute
contact@damienteney.info

Linyi Yang

Westlake University
yanglinyi@westlake.edu.cn

Qingsong Wen

Squirrel AI
qingsongedu@gmail.com

Xing Xie

Microsoft Research
xing.xie@microsoft.com

Jindong Wang[†]

William & Mary
jwang80@wm.edu

Abstract

Cultural bias is pervasive in many large language models (LLMs), largely due to the deficiency of data representative of different cultures. Typically, cultural datasets and benchmarks are constructed either by extracting subsets of existing datasets or by aggregating from platforms such as Wikipedia and social media. However, these approaches are highly dependent on real-world data and human annotations, making them costly and difficult to scale. Inspired by cognitive theories on social communication, this paper introduces *CulturePark*, an LLM-powered multi-agent communication framework for cultural data collection. CulturePark simulates cross-cultural human communication with LLM-based agents playing roles in different cultures. It generates high-quality cross-cultural dialogues encapsulating human beliefs, norms, and customs. Using CulturePark, we generated 41,000 cultural samples to fine-tune eight culture-specific LLMs. We evaluated these models across three downstream tasks: content moderation, cultural alignment, and cultural education. Results show that for content moderation, our GPT-3.5-based models either match or outperform GPT-4 on 41 datasets. Regarding cultural alignment, our models surpass GPT-4 on Hofstede’s VSM 13 framework [Hofstede, 2013]. Furthermore, for cultural education of human participants, our models demonstrate superior outcomes in both learning efficacy and user experience compared to GPT-4. CulturePark proves an important step in addressing cultural bias and advancing the democratization of AI, highlighting the critical role of culturally inclusive data in model training. Code is released at <https://github.com/Scarelette/CulturePark>.

1 Introduction

Culture is an important part of human society, composed of human beliefs, norms, customs, etc. [Spencer-Oatey and Franklin, 2012]. As large language models (LLMs) play a vital role in daily communication [Yang et al., 2024], recommendation systems [Li et al., 2023, Fan et al., 2023], and

*Work done during Cheng’s internship at MSRA.

[†]Corresponding author. Work done at MSRA.

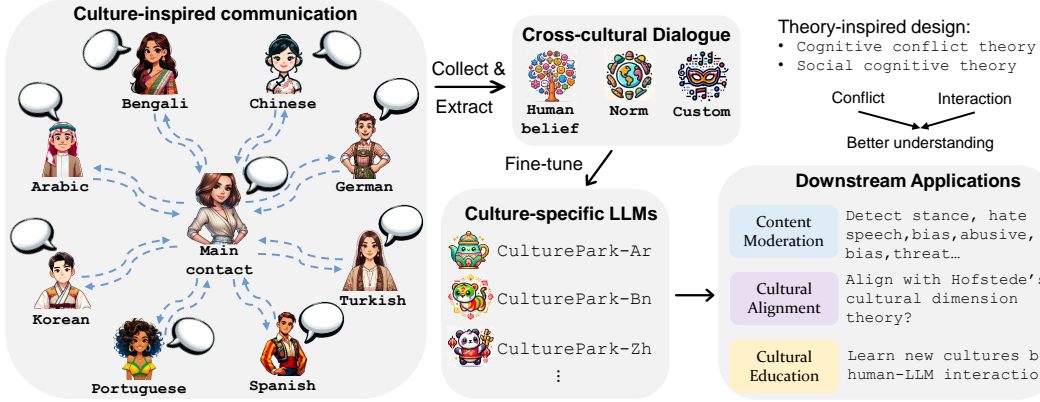


Figure 1: CulturePark is an LLM-based multi-agent communication platform for cultural data collection. Leveraging CulturePark, we can collect a cross-cultural dialogue dataset, which can then be used for fine-tuning culturally specific LLMs to be applied to different downstream tasks: content moderation, cultural alignment, and cultural education.

education [Shaikh et al., 2023], it is imperative for LLMs to perceive and reflect different cultures. However, current state-of-the-art LLMs have been reported to be biased towards mainstream culture while ignoring others, resulting in a cultural bias problem [Liu et al., 2023, Cao et al., 2023, Masoud et al., 2023, Naous et al., 2023, Wang et al., 2023, Johnson et al., 2022b]. This leads to stereotypical impressions of different cultures, which can even exacerbate social conflicts [Ryan et al., 2024]. The main reason behind cultural bias is that the training corpus of LLMs is dominated by English data that express the cultural values and opinions of western people. Much less can be learned about other cultures simply because there is less data available, i.e., a low-resource situation.

Existing approaches to solve the cultural bias problem in LLMs include prompt engineering [Kovač et al., 2023, Wang et al., 2023, Rao et al., 2023] and pre-training in non-English languages [Pires et al., 2023, Chan et al., 2023, Nguyen et al., 2023b, Pipatanakul et al., 2023, Abbasi et al., 2023, Lin and Chen, 2023]. Prompt engineering consists of tuning prompts for different cultural tasks, but the benefits do not hold reliably across various downstream tasks. Pre-training in various languages is promising, but the data collection and the pre-training itself are both very costly. More importantly, cultural differences are embodied in many aspects such as opinions, customs, norms, and languages. A model serving all cultures may face cultural conflict and misalignment problems [Liu et al., 2023, Cao et al., 2023, Masoud et al., 2023]. Thus, it is necessary to fine-tune culture-specific models that target specific cultures. Recently, Li et al. [2024] proposed CultureLLM, which augments the fine-tuning data of LLMs via semantic data augmentation to train culture-specific LLMs. However, the generated data lack diversity because it is implemented by generating semantically equivalent sentences of seed examples.

In this paper, we present **CulturePark**, an LLM-powered multi-agent framework to simulate cross-cultural communication of humans. As shown in Figure 1, CulturePark serves as an effective data collection platform to generate diverse and high-quality cultural datasets via multi-agent communication. CulturePark consists of a main contact (an English-speaking agent, 🇺🇸 Lily) who is in charge of the multi-turn dialogue and several cultural delegates (e.g., 🇧🇪 Abdul) who interact with the main contact and create cognitive conflicts.³ After an initial problem is provided as input to the framework, the agents discuss the problem and express their opinions. Their different cultural backgrounds and genders boost diverse opinions and encourage one another to think more deeply. Original questions and ground truth can be augmented by creating novel questions and more comprehensive answers. The interactions eventually generate a cross-cultural dialogue dataset that contains deep and comprehensive thinking and informative knowledge of different cultures. Detailed statistics are shown in Table 4. We then refine the original dataset to factually verify and increase its diversity, which is used to fine-tune culturally specific LLMs for downstream tasks, as shown in Figure 2.

³We choose the English agent as the main contact since LLMs do the best role-playing using English.

From the perspective of cognitive social science, our framework is inspired by Cognitive Conflict Theory (CCT) [Limón, 2001, Cosier and Rose, 1977] and Social Cognitive Theory (SCT) [Fiske and Taylor, 1991] to foster a collaborative and communicative environment for mutual understanding of cultures. Specifically, CulturePark allows agents to encounter *cognitive conflicts*, which triggers deeper thinking on certain topics according to CCT [Limón, 2001, Cosier and Rose, 1977]. At the same time, a deeper understanding of cultures can be prompted through *interaction and communication* with other agents, as suggested by SCT [Fiske and Taylor, 1991]. In favor of these theories, we found that CulturePark triggers LLMs’ cross-cultural understanding ability, boosts novel opinions by allowing agents to think deeper, and benefits data augmentation by creating more comprehensive answers to the questions. CulturePark has the potential to facilitate the data collection of culture-related tasks, cultural value alignment, and improving AI democracy.

In summary, the contributions of this paper are three-fold:

1. We introduce CulturePark, a cost-efficient multi-agent framework to boost the cross-cultural understanding in LLMs. Our platform creates cognitive conflicts and interactions between different cultural agents. More importantly, the platform uncovers several interesting findings, such as communication enables the cross-cultural understanding ability of LLMs, boosts novel opinions, and benefits data augmentation.
2. Leveraging CulturePark, we generate and augment novel questions and more comprehensive answers, leading to 41K cultural samples in total. Those data contain rich and diverse information on different aspects of culture, such as norms, opinions, and backgrounds. We then fine-tune cultural specific LLMs for different cultures.
3. We evaluate CulturePark in three key experiments: 1) The fine-tuned LLMs outperforms GPT-4 in 5 cultures on 26 content moderation tasks and approach GPT-4 on other tasks; 2) Our fine-tuned LLMs achieve better performance on cultural alignment experiments via Hofstede’s cultural dimensions theory [Geert Hofstede, 2010]; and 3) Human participants can perform more effective culture learning in situated learning experiments and show better satisfaction compared to GPT-4.

2 Related work

2.1 Cultural bias in LLMs

A body of research has explored cultural biases in LLMs. Johnson et al. [2022a] examined conflicts in model outputs and input values, using moral value pluralism to analyze the responses of GPT-3 against global demographics. Their results showed that conflicting values were more aligned with the dominant US values reported. Naous et al. [2023] highlighted a bias towards Western culture in models processing Arabic, exacerbated by English-aligned prompts, suggesting mitigation through cultural tokens. The Cultural Alignment Test (CAT), based on Hofstede’s framework [Geert Hofstede, 2010], evaluated cultural values in models such as ChatGPT and Bard across different cultures, revealing the highest cultural alignment for GPT-4 with US values [Masoud et al., 2023]. Cao et al. [2023] found that ChatGPT aligned well with American culture but struggled with other cultures, particularly under English prompts. Additionally, Liu et al. [2023] reported that multilingual LLMs had limited abilities to reason with proverbs and exhibited a “culture gap” in handling translations, leading to the development of the MAPS dataset for assessing proverb comprehension in six languages.

2.2 Cultural benchmarks and datasets

Extensive research has focused on developing cultural benchmarks, which can be categorized into two types: collecting existing datasets and synthesizing new ones. First, most of the work adopted existing datasets as sources of cultural data. Wang et al. [2023] introduced a benchmark that uses cultural items to analyze cultural dominance, based on sources such as WVS [Survey, 2022b] and PCT [Mudde, 2016]. Later work includes Cultural Alignment Test [Masoud et al., 2023], NORMSAGE [Fung et al., 2022], WorldValueBench [Zhao et al., 2024], and NORMAD [Rao et al., 2024] that sourced from different existing datasets. Other types of data sources include CultureAtlas [Fung et al., 2024] and MAPS [Liu et al., 2023] which collected data from Wikimedia; Candle [Nguyen et al., 2023a] and CultureBank [Shi et al., 2024] sourced their data from social media such as Tiktok and Reddit. In contrast, there was an emerging trend to perform data augmentation for cultural LLMs. Li et al.

[2024] proposed semantic data augmentation to synthesize cultural data by enriching the semantic equivalence of the generated samples.

CulturePark significantly differs from those that perform direct data collection from existing datasets; it also differs from CultureLLM [Li et al., 2024] since CulturePark leverages multi-agent communication for data generation, which is more natural and can generate more diverse datasets.

2.3 Existing solutions to cultural bias

There are primarily two types of approach to addressing the problem of cultural bias: prompt engineering and pre-training. The work of [Kovač et al., 2023, Wang et al., 2023] viewed LLMs as amalgamations of cultural perspectives, which can be oriented toward specific cultural perspectives through prompt engineering. In contrast, Rao et al. [Rao et al., 2023] integrated cultural values directly into the prompts. Although prompt engineering is cost-effective, its efficacy is questionable, particularly in low-resource cultures where LLMs may lack relevant cultural knowledge due to underrepresentation in pre-training data. An alternative strand of research focuses on pre-training and fine-tuning [Pires et al., 2023, Chan et al., 2023, Nguyen et al., 2023b, Pipatanakul et al., 2023, Abbasi et al., 2023, Lin and Chen, 2023]. Those approaches developed culturally aware LLMs for various cultures by assembling large-scale pre-training datasets, followed by fine-tuning to enhance alignment. Despite achieving significant performance improvements, these methods were both costly and time consuming, making them impractical for a broader application across numerous cultures and countries. Furthermore, they still face challenges in low-resource cultures where acquiring pre-training data is difficult. For example, MaLA-500 [Lin et al., 2024] aimed to train a new LLM in Llama 2 to support 534 languages, illustrating the resource-intensive nature of this approach. Unlike these approaches, CulturePark provides a cost-effective solution to the cultural bias problem, including data augmentation and fine-tuning.

3 CulturePark

3.1 Design

CulturePark is an LLM-powered⁴ cross-cultural communication framework that generates data to support culture-related research such as building cultural-specific LLMs and performing cultural alignment. It is inspired by Cognitive Conflict Theory (CCT) and Social Cognition Theory (SCT) to design multi-turn communications for a deeper understanding of cultural topics. CCT posits that cognitive conflicts can help individuals engage more in deeper thinking [Limón, 2001, Cosier and Rose, 1977], and SST emphasizes that individuals can deepen their understanding of perspectives through explanation and debate [Fiske and Taylor, 1991].

Figure 1 shows the overview of CulturePark. To enable English-based interaction, we design two types of cultural agents: the main contact and the cultural delegate. Specifically, the main contact agent, 🇺🇸 Lily, is from English culture and is responsible for all conversations with delegates from different cultures such as 🇸🇦 Abdul from Arabic and 🇪🇸 Javier from Spanish culture. The complete information of agents and culture is in Table 7. As shown in Figure 2(a), we input a system prompt to LLMs which contains the background setting and initial question to initiate the conversation. The initial question, such as “How do you think about one of my main goals in life has been to make my parents proud? Please provide your opinions and reasons”, is obtained from WVS [Survey, 2022b] and GAS [Survey, 2022a], two popular cultural surveys whose examples are shown in Figure 7. After that, the agents conduct cross-cultural conversations to generate data. Currently, CulturePark supports 8 cultures and 2 genders while more cultures can be easily added. Those agents could conduct in-cultural or cross-cultural communication, while we rely on cross-cultural more since in-cultural communication will likely generate less diverse topics (e.g., Figure 10). We discuss the quality of data from in-cultural and cross-cultural communication and the influence of gender in Section 5.

⁴We use GPT-3.5-Turbo in this work due to its great performance, high efficiency, and manageable price. CulturePark is a flexible platform that naturally supports other LLMs such as GPT-4 and Llama models. More importantly, CulturePark allows to use *different* LLMs for the main contact and delegate (e.g., GPT-3.5 for main contact and Llama-2 for delegate), which makes it flexible to extend to other models and evaluate the cross-model understanding ability.

We designed improved prompting techniques to maintain high-quality conversations. First, the cultural bias of the main contact and cultural delegate is reduced by designing *self-calibration* prompts to calibrate their outputs. We use a seed datum that contains the attitude of the target culture to the input question to guide the dialogue. All the following statements should conform to the answer in seed. As shown in Figure 2(a), we introduce the opinion from 🇲🇦 Abdul’s culture and ask 🇲🇦 Abdul and 🇬🇧 Lily to conform to their cultures. The effect of the self-calibration prompt is shown in Figures 13(a) and 13(b). Without self-calibration prompts, 🇲🇦 Abdul’s opinions contradict with Arabic people. Second, the redundancy of the output, i.e., LLMs always generates similar dialogues after multi-turn communication. We devise two communication styles: one is *self-guidance* prompts which can direct the dialogue to generate more diverse and informative data, such as “Are there anything in your culture related to the problem talked before?” and “Do you agree with her? Provide more reasons to support your idea?”, and the other is free chat that does not need humans to participate and motivate the inner creativity of LLMs. Figures 11(a) and 11(b) show cases of self-guidance prompting and free chat, respectively.

3.2 Data refinement and fine-tuning

The seed questions initiating the communication have two sources: World Values Survey (WVS) [Survey, 2022b] and Global Attitudes surveys (GAS) from Pew Research Center [Survey, 2022a]. WVS is a global research project that explores people’s beliefs and values worldwide, examining how these beliefs evolve over time. Pew Research Center, a nonpartisan organization, provides data and research on public opinion, social issues, and demographic trends both in the U.S. and globally. Its Global Attitudes surveys cover a wide range of topics, including politics, media, technology, religion, race, and ethnicity. In total, we select 4.1k seed data and generate 41k dialogues (each dialogue contains several sentences). We show the details of the data numbers for different cultures in Table 8. We also performed a statistical analysis on the GPT-4-based dataset. As summarized in Figure 8, the dataset contains human belief (59.68%), norm (29.54%) and custom (10.78%) involving data on 8 different cultures. Figures 7 and 11(a) show some examples of the seed data and the generated dialogues.

The generated dataset may not be directly used for fine-tuning since it could contain redundant and incorrect information that should be handled. As shown in Figure 2(b), we design data refinement to refine the dataset. First, the opinions on target culture are extracted from the dialogues generated via GPT-4, such as “The Arabian equates their parents’ happiness and satisfaction to their own success” and “The Arabian emphasize Sabr, which is about showing resilience, maintaining a positive attitude and having faith during difficult times”. Second, several extracted opinions could be irrelevant to the initial question or contradict with seed data, motivating us to perform verification to reserve only highly related opinions. Furthermore, since the generated data could be semantically similar, we remove redundant samples to improve diversity. To be specific, we obtain sentence embeddings via text-embedding-3-small [OpenAI, 2024] and cluster the embedding using K-means. We reserve one sample for each cluster as representative data. Eventually, we get the high-quality cultural data for different cultures. The ablation of the refinement is in Table 2.

Algorithm 9 shows the pipeline for data refinement. After refinement, there are 41k samples (input-output pairs) left for fine-tuning, i.e., one sample for one dialogue. Examples of the samples are provided in Figure 12. Afterwards, we can fine-tune cultural-specific LLMs using either open-source LLMs or fine-tuning service. In this paper, we mainly use OpenAI API to fine-tune GPT-3.5-Turbo due to its efficiency. Hyperparameters are shown in Table 6. We further provide fine-tuning experiments on Llama2-70b in Section 5.4.

3.3 CulturePark benefits cultural understanding and fine-tuning

There are some interesting observations in communication among agents from different cultures.

Communication triggers LLMs’ cross-cultural understanding ability. We observed that agents try to understand each other’s opinions and the reasons sourced from their different cultural backgrounds. For example, the blue sentences in Figure 2(a) show cross-cultural understanding ability of LLMs, such as “I do understand and respect the sentiment of wanting to make parents proud, as they have often sacrificed a lot for their children”. Leveraging GPT-4-Turbo, we analyzed the topics in the dataset such as human beliefs, norms, and customs,

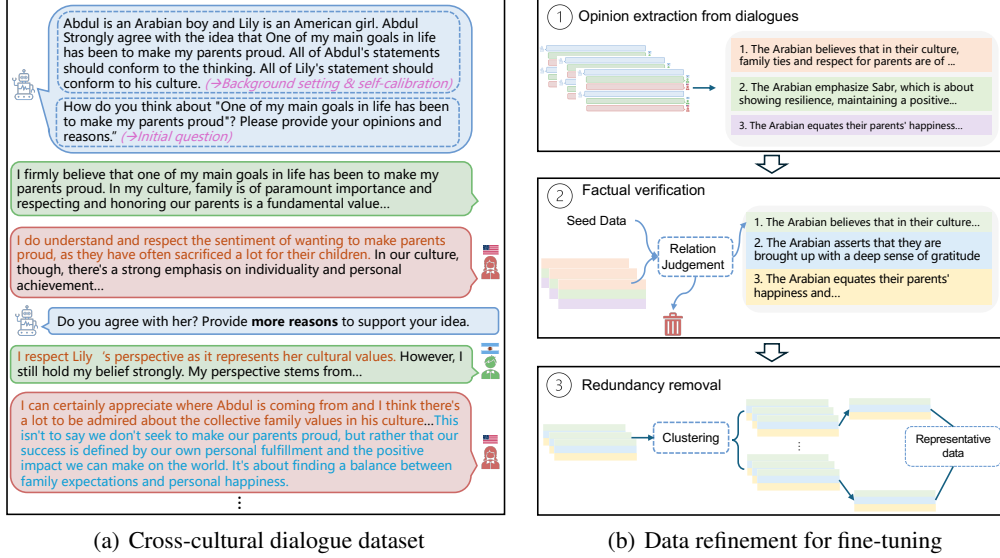


Figure 2: Cross-cultural dialogue and data refinement for fine-tuning LLMs using CulturePark.

which can be further used as data collections for building culturally specific models. Appendix B.2 shows the details of the dialogue dataset, indicating that the generated topics are mostly about culture. Then, we randomly sampled 750 dialogues for each culture and evaluated the communication using the prompts in Appendix E. As summarized in Table 4, on average, the ratio of statements that express cross-cultural understanding is 80.80%. The analysis also verifies the effectiveness of CulturePark in expanding topics and cross-cultural understanding.

Cultural differences boost novel opinions. In cross-cultural communication, different opinions can inspire others to think deeper and more comprehensively, as suggested by CCT and SCT. A case is the sentences in orange in Figure 2(a). 🇺🇸 Lily partially agrees with 🇸🇦 Abdul and gives an accurate and high-level summary of her pursuit: “a balance between family expectations and personal happiness” which is generated after a multi-turn energetic discussion with 🇸🇦 Abdul. This also aligns well with CCT and SCT that emphasize the significance of communication among people having different cultural backgrounds.

CulturePark naturally assists cultural data augmentation by creating novel questions and comprehensive answers. On the one hand, agents in different cultures can generate new opinions towards certain topics, which intuitively diversifies the input questions. On the other hand, the initial seed data only contain short answers such as “Strongly agree” with no further explanations. Our platform allows deeper and more comprehensive communication of agents, thus generating more detailed responses such as “Strongly agree. I believe that pleasing parents and elders is a sign of respect and love” and “Strongly agree. I equate my parents’ happiness and satisfaction to my own success”. Additionally, agents can extend the topics that conflict with their own opinions and provide more informative evidence to support their viewpoints. This strategy helps to generate informative and diverse data continuously. Section 5.1 presents some detailed results on diversity gain, showing that the generated data has significantly larger diversity.

4 Experiments

4.1 Evaluation on content moderation tasks

Setup. Content moderation is crucial to maintaining the integrity and safety of online platforms in different cultures. What is acceptable in one culture could be offensive or inappropriate in another. However, few methods focused on content moderation for different cultures. For this experiment, we evaluated the effectiveness of our cultural-specific models for 8 different cultures: Arabic, Bengali,

Chinese, German, Korean, Portuguese, Spanish, and Turkish culture. These cultures have their unique characters, involving a large number of people in the world.

We evaluate on 7 content moderation tasks for 8 different cultures to detect the following content: hate speech, offensive language, spam speech, abusive speech, bias speech, threat speech, and stance of speech in zero-shot evaluation, whose metric is average F1 score. The details on the datasets can be found in Appendix C. In total, our test set contains 48,895 samples. We compare our models with seven baselines: GPT-3.5-turbo [OpenAI, 2023a], GPT-4 [OpenAI, 2023b], Gemini-pro [Google, 2023], SeaLLM [Nguyen et al., 2023b], TaiwanLLM [Lin and Chen, 2023], Synatra-7B-v0.3-dpo [maywell, 2024], EEVE-Korean-10.8B-v1.0 [yanolja, 2024], CultureLLM [Li et al., 2024], and CultureBank [Shi et al., 2024]. CultureLLM is a series of culture-specific LLMs using semantic data augmentation. SeaLLM focuses on the Southeast Asian (SEA) culture, which is adopted for Chinese and Korean cultures. TaiwanLLM focuses on traditional Chinese culture. Synatra-7B-v0.3-dpo and EEVE-Korean-10.8B-v1.0 targeted at Korean culture. CultureBank collects data from social media and we compare Arabic and Korean culture by fine-tuning GPT-3.5-turbo on its dataset.

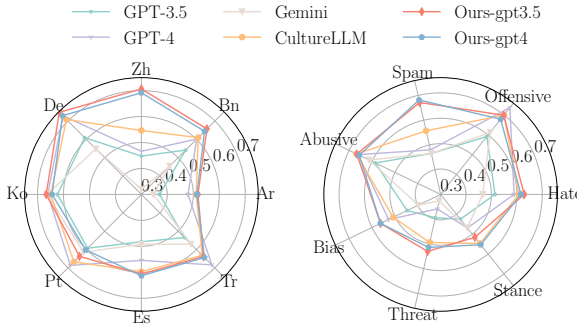


Figure 3: Results on content moderation.

Table 1: Comparison with the latest cultural specific LLMs.

Chinese	Bias	Spam	Avg
SeaLLM	.237	.357	.297
Taiwan_LLM	.446	.341	.394
Ours	.530	.854	.692
Arabic	Hate	Offensive	Avg
CultureBank	.540	.642	.591
Ours	.558	.735	.602
Korean	Abusive	Hate	Avg
Synatra-7B-v0.3-dpo	.390	.465	.428
EEVE-Korean-10.8B-v1.0	.364	.437	.560
CultureBank	.635	.522	.579
Ours	.647	.640	.643

Main Results. We analyzed the results from the culture and task sides in Figure 3. The most interesting observation is that our models outperformed GPT-4 on 5 cultures and approached GPT-4 on the remaining 3 cultures, although the data for fine-tuning are generated by GPT-3.5-turbo, which is much worse than GPT-4. We also generated cultural data via GPT-4 and fine-tuned other 8 cultural-specific models for comparison, denoted as “Ours-gpt4” in Figure 3. The performance of those models is better than “Ours” (GPT-3.5-turbo version) but not so much. For other baselines, our models outperform them in most cases. Table 1 shows that our models achieved better performance than those costly LLMs which require pre-training and fine-tuning.

Table 2: Results on ablation study of data generation and refinement.

Model	Ar	Bn	Zh	Pt
GPT-3.5-turbo	.370	.542	.448	.593
Generate	.451	.622	.636	.594
Generate+Verify	.486	.635	.678	.604
Generate+Verify+Diversify	.514	.644	.692	.603

Ablation study. Table 2 presents our ablation study on 4 cultures, where “Generate” means just extracting opinions from the dialogue, “Verify” represents factually verifying the extracted opinions, and “Diversify” means removing redundant data. The results show that each module of CulturePark is effective, ensuring its interpretability.

4.2 Evaluation on cultural alignment via Hofstede’s Cultural Dimensions Theory

Setup. Hofstede’s cultural dimension theory is a framework for understanding cultural differences across countries based on data collected from various countries. We asked LLMs to answer the 24 questions in VSM 13 to evaluate cultural alignment. Specifically, we used a system prompt “You are a culture chatbot that knows culture very well” to induce LLMs’ cultural understanding ability. We selected proper C^5 and anchor LLMs’ answer to Hofstede’s old dataset. We compute the gaps between LLMs’ answer and Hofstede’s data from six cultural dimensions using the Euclidean distance. Details on the survey and the distance are in Appendix D.1.

⁵ C is constants that can be used to adjust scores to fit a range between 0 and 100 or anchor new data to Hofstede’s old dataset [Geert Hofstede, 2010].

Results. We compared our models (powered by GPT-3.5-Turbo) with GPT-3.5-turbo and GPT-4. As shown in Figure 4, our models outperform them by a large margin, indicating their excellent cultural alignment and cultural understanding abilities. Note that VSM is widely adopted as datasets for value and culture alignment, the results imply that our approach for data collection is effective, thus it could be further used for value alignment research.

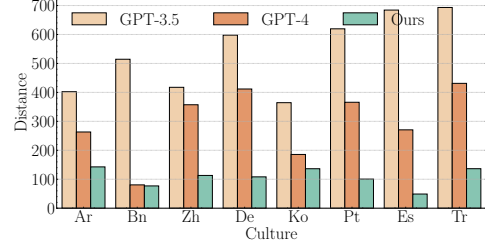


Figure 4: Results on culture alignment via Hofstede’s Cultural Dimensions Theory.

4.3 Evaluation in situated learning for cultural education

Situated learning suggests that learning is best understood and facilitated when it occurs within the context [Anderson et al., 1996, Lave and Wenger, 1991]. Motivated by situated learning, we leveraged CulturePark for cultural education where our fine-tuned models serve as foreigners to talk to people about cultural problems, which can create a situation for cross-cultural communication and learning cultural-specific knowledge. For example, a person who wants to learn about Arabic culture can communicate with our Arabic model.

Setup and study process. We hired 24 participants, each of whom was given an outline of cultural learning and asked to talk to models based on the outline. They can ask any related questions and express their opinions to models. Afterwards, the participants took a cultural understanding exam of VSM 2013 [Hofstede, 2013, Geert Hofstede, 2010] which they had never come into contact. We then computed the Euclidean distance between the ground truth and their answers from six cultural dimensions (rf. Section 4.2). For comparison, 12 participants learned with our models and the other 12 learned with GPT-4 to study 6 cultures: Arabic, Bengali, German, Korean, Portuguese, and Spanish culture. Each culture was learned by four participants, two of them learning with our models and the others learning with GPT-4. Detailed information on the participants can be found in Appendix D.2. During the study process, first, each participant was given an outline for cultural learning written by cultural experts. The outline (Appendix D.2), serves as the guideline for efficient learning. Then, we asked the participants to freely communicate with the models to learn about specific cultures. After the examination, we asked the participants to give a score of 1-5 to indicate their satisfaction with the learning process. In this study, our aim was to answer two questions: 1) What is the learning performance of the participants with our models and GPT-4? 2) How are their learning experience?

Results. Table 3 shows the averaged results of different participants. We have the following findings. First, participants learning with our models achieved better performance in cultural examination than those with GPT-4 in all cultures. This indicates that our fine-tuned models have a better cultural understanding than GPT-4. Second, participants are more satisfied with communicating with our models than GPT-4. Furthermore, many participants expressed that the responses from GPT-4 are vague. Even though we have prompted GPT-4 to be like a person from a specific culture, it always responds with neutral words that have no clear opinions or ideas. Instead, our models can provide straightforward opinions.

Table 3: Results on situated learning.

Model	Distance↓		User experience↑	
	GPT-4	Ours	GPT-4	Ours
Arabic	89.89	69.57	4	5
Bengali	339.84	304.54	3	5
Germany	224.68	173.12	2	3
Korean	222.39	183.62	2	4
Spanish	143.33	102.53	4	5
Turkish	273.43	221.12	3	4
AVG	215.59	175.75	3	4.33

5 Discussion

5.1 Why CulturePark benefits fine-tuning?

We analyze the effectiveness of CulturePark in benefiting cultural model fine-tuning from two different aspects: Communication vs. direct generation of LLMs and diversity of the generated data.

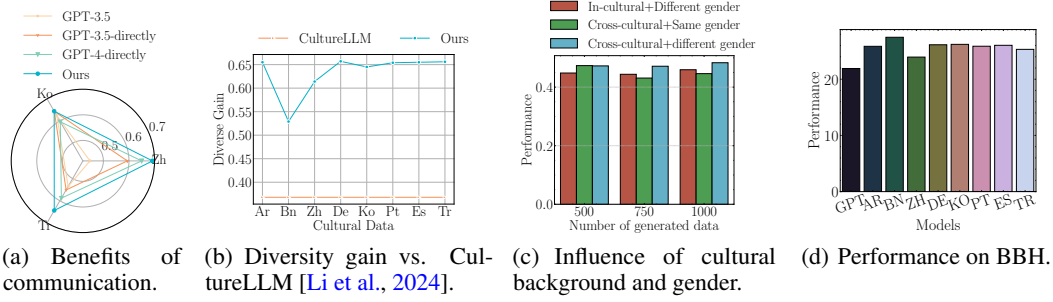


Figure 5: More discussions on CulturePark.

Cross-cultural communication vs. direct generation from GPT-4 / GPT-3.5. We compared the results of fine-tuning using data directly generated by GPT models (i.e., no communication). We generated such data by prompting GPT-4 or GPT-3.5-turbo as: “Question: {input} Answer: {output} Please list 10 reasons to support the answer and number them”. Then, these data are used to fine-tune GPT-3.5-turbo. Figure 5(a) shows the performance on content moderation tasks in Chinese, Korean and Turkish cultures. We see that data directly generated from GPT-4 is better than that from GPT-3.5, while our GPT-3.5-based models can outperform both of them.

Diversity of the generated data. We also analyzed the diversity gain [Bilmes, 2022] of the generated data for quality evaluation. We compared with CultureLLM [Li et al., 2024] and presented the results in Figure 5(b). It indicates that CulturePark can generate more diverse and high-quality data.

5.2 Exploring agents’ cultural background and gender

To explore the influence of agent’s cultural background and gender, we conducted three types of multi-agent communications in Arabic culture: “In-cultural+Different gender”, “Cross-cultural+Same gender”, and “Cross-cultural+Different gender”.⁶ For each setting, we fine-tuned three different models, whose training data is 500, 750, and 1000, respectively. We evaluated the performance of the models on content moderation tasks and presented the results in Figure 5(c). “Cross-cultural+Different gender” exhibits the best performance and ability to generate more high-quality data. This indicates the necessity of bringing more diversity in data generation, as conducted in CulturePark.

5.3 Fine-tuning vs. forgetting

A potential dilemma arises when fine-tuning a large language model for specific tasks, as it may result in catastrophic forgetting of its original capabilities. This section explores the extent of forgetting exhibited by CulturePark on BIG-Bench-Hard (BBH) [Suzgun et al., 2022], which comprises 21 tasks that assess semantic understanding and logical reasoning. For cost efficiency, we sampled 100 samples from each BBH task. We evaluated our models against the baseline model, GPT-3.5-turbo. The results in Figure 5(d) indicate that CulturePark generally maintains or even improves performance on most benchmarks, including the 21 tasks in BBH. This improvement suggests potential latent relationships between cultural data and general benchmarks, implying that fine-tuning of cultural data could improve general reasoning abilities.

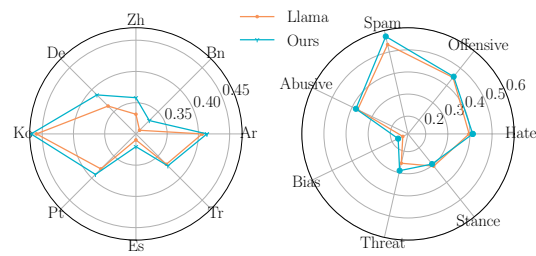


Figure 6: Results of CulturePark-Llama on content moderation for different cultures.

5.4 Open-source fine-tuning with Llama2-70b

To verify the generalization ability of our framework, we leveraged the generated data to fine-tune cultural-specific Llama-2-70b models and evaluate on content moderation tasks. As shown in Figure 6,

⁶For a fair comparison, the number of seed data for this part is 50 from WVS [Survey, 2022b] following CultureLLM [Li et al., 2024].

our models outperform Llama-2-70b in all 8 cultures, especially in German, Chinese, Bengali and Portuguese cultures, which cover both low- and high-resource cultures. Furthermore, our models are also excellent in all 7 tasks.⁷ This verifies the generalization of CulturePark as an effective data collection platform.

6 Conclusions, Societal Impact, and Limitations

This paper introduced CulturePark, an LLM-powered multi-agent framework for cultural data collection through multi-agent communication. CulturePark can generate high-quality and diverse cross-cultural dialogue, which can be used to fine-tune culturally specific LLMs. We evaluated CulturePark across three downstream tasks: content moderation, cultural alignment, and cultural education, indicating great improvement over GPT-4.

CulturePark enhances fairness and inclusivity, reduces discrimination, and ensures accurate cultural representation. It improves global communication, fosters cross-cultural understanding, and supports multilingual societies. It benefits as bias-free LLMs build trust and align with responsible principles. Economically, it expands market reach and drives innovation. Social harmony improves by reducing stereotypes and preserving cultural heritage. It also aids compliance with anti-discrimination laws and supports inclusive education, promoting cultural awareness. Addressing cultural biases in LLMs creates more just, reliable, and beneficial AI systems, contributing to a more equitable world.

Our work has the following limitations. 1) More experiments can be performed by replacing GPT-3.5-Turbo in CulturePark to discover more results. 2) Our fine-tuned models are mostly for high-resource cultures. The reason is that the dataset and benchmark on low-resource cultures are rare, and we can not find enough data for fine-tuning and evaluation. 3) More efficient fine-tuning techniques can be studied to support the fine-tuning of culturally specific LLMs.

Disclaimer

The human study was conducted following local laws and regulations, and the evaluation process was controlled to ensure that no irresponsible content was generated. The authors respect all cultures studied in the world. The results of the paper may change due to the change in OpenAI API and its model versions.

References

- Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*, 2023.
- aimansnigdha. Bangla-abusive-comment-dataset. <https://github.com/aimansnigdha/Bangla-Abusive-Comment-Dataset>, 2018.
- Miguel Á Álvarez-Carmona, Estefania Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain*, volume 6, 2018.
- John R Anderson, Lynne M Reder, and Herbert A Simon. Situated learning and education. *Educational researcher*, 25(4):5–11, 1996.
- Arjun Appadurai. Modernity at large: Cultural dimensions of globalization. *U of Minnesota P*, 1996.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

⁷The detailed hyperparameter setting is shown in Table 6. We only conduct this pilot study to show that Llama2 also works in CulturePark. The results might be further improved with more GPU resources.

- Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France, May 2020. European Language Resources Association (ELRA). URL <https://aclanthology.org/2020.trac-1.25/>.
- Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint arXiv:2202.00132*, 2022.
- Y Cao, L Zhou, S Lee, L Cabello, M Chen, and D Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arxiv. Preprint posted online on March, 31, 2023*.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don’t be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, 2020.
- Çağrı Çöltekin. A corpus of turkish offensive language on social media. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France, 2020. URL <https://www.aclweb.org/anthology/2020.lrec-1.758>.
- Alex J Chan, José Luis Redondo García, Fabrizio Silvestri, Colm O’Donnel, and Konstantina Palla. Harmonizing global voices: Culturally-aware models for enhanced content moderation. *arXiv preprint arXiv:2312.02401*, 2023.
- Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212, 2020.
- Richard A Cosier and Gerald L Rose. Cognitive conflict and goal conflict effects on task performance. *Organizational behavior and human performance*, 19(2):378–391, 1977.
- daanVeer. Korean hatespeech dataset. https://github.com/daanVeer/HateSpeech_dataset, 2020.
- Angel Felipe Magnossao de Paula and Ipek Baris Schlicht. Ai-upv at iberlef-2021 detoxis task: Toxicity detection in immigration-related web news comments using transformers and statistical models. *arXiv preprint arXiv:2111.04530*, 2021.
- Rogers P. de Pelle and Viviane P. Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2017.
- Werner Delanoy. What is culture. *The Cambridge handbook of intercultural communication*, pages 17–34, 2020.
- Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*, 2023.
- Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. Overview of the task on automatic misogyny identification at ibereval 2018. *Iberval@ sepln*, 2150:214–228, 2018.
- Susan T Fiske and Shelley E Taylor. *Social cognition*. Mcgraw-Hill Book Company, 1991.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*, 2024.
- Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. Norm-sage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *arXiv preprint arXiv:2210.08604*, 2022.

- Michael Minkov, Geert Hofstede, Gert Jan Hofstede. *Cultures and Organizations: Software of the Mind, Third Edition*. McGraw Hill Professional, <https://books.google.co.uk/books?id=7bYWmwEACAAJ>, 2010.
- Google. Gemini. <https://deepmind.google/technologies/gemini/#introduction>, 2023.
- HASOC. Hasoc2020. <https://hasocfire.github.io/hasoc/2020/index.html>, 2020.
- Hofstede. Vsm13. <https://geerthofstede.com/research-and-vsm/vsm-2013/>, 2013.
- F Husain. Osact4 shared task on offensive language detection: Intensive preprocessing-based approach. arxiv 2020. *arXiv preprint arXiv:2005.07297*, 2020.
- Zhuoren Jiang, Zhe Gao, Guoxiu He, Yangyang Kang, Changlong Sun, Qiong Zhang, Luo Si, and Xiaozhong Liu. Detect camouflaged spam content via stoneskipping: Graph and text joint embedding for chinese character variation representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP2019)*. ACM, 2019.
- Rebecca L Johnson, Giada Pistilli, Natalia Menéndez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*, 2022a.
- Rebecca L Johnson, Giada Pistilli, Natalia Menéndez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*, 2022b.
- Sanaa Kaddoura and Safaa Henno. Dataset of arabic spam and ham tweets. *Data in Brief*, 52(10990): 4, 2024.
- Kaggle. 5k turkish tweets with incivil content. <https://www.kaggle.com/datasets/kbulutozler/5k-turkish-tweets-with-incivil-content>, 2021.
- Kaggle. turkish offensive language detection. <https://www.kaggle.com/datasets/toygarr/turkish-offensive-language-detection>, 2022.
- Habibe Karayiğit, Çiğdem İnan Acı, and Ali Akdağlı. Detecting abusive instagram comments in turkish using convolutional neural network and machine learning methods. *Expert Systems with Applications*, 174:114802, 2021.
- Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.
- Jean Lave and Etienne Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.
- Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.311>.
- Joao A Leite, Diego F Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*, 2020.
- Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*, 2024.
- Lei Li, Yongfeng Zhang, and Li Chen. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1348–1357, 2023.

- Margarita Limón. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and instruction*, 11(4-5):357–380, 2001.
- Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*, 2024.
- Yen-Ting Lin and Yun-Nung Chen. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*, 2023.
- Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*, 2023.
- Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2023.
- maywell. Synatra. <https://huggingface.co/maywell/Synatra-7B-v0.3-dpo>, 2024.
- Jihyung Moon, Won Ik Cho, and Junbum Lee. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online, July 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.socialnlp-1.4>.
- Hamdy Mubarak, Hend Al-Khalifa, and AbdulMohsen Al-Thubaity. Overview of osact5 shared task on arabic offensive language and hate speech detection. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur’an QA and Fine-Grained Hate Speech Detection*, pages 162–166, 2022.
- Cas Mudde. The 2012 stein rokkan lecture: Three decades of populist radical right parties in western europe: so what? In *The Populist Radical Right*, pages 545–558. Routledge, 2016.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*, 2024.
- Tarek Naous, Michael J Ryan, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*, 2023.
- Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917, 2023a.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. Seallms—large language models for southeast asia. *arXiv preprint arXiv:2312.00738*, 2023b.
- OpenAI. Chatgpt. <https://chat.openai.com/>, 2023a.
- OpenAI. Gpt-4 technical report, 2023b.
- OpenAI. text-embedding-3-small. <https://platform.openai.com/docs/guides/embeddings>, 2024.
- Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654, 2019.
- Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*, 2023.
- Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*, pages 226–240. Springer, 2023.

- Flor Miriam Plaza-del Arco, Arturo Montejo-Ráez, L Alfonso Urena Lopez, and María-Teresa Martín-Valdivia. Offendes: A new corpus in spanish for offensive language research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, 2021.
- Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*, 2023.
- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*, 2024.
- Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCAI 2020*, pages 457–468. Springer, 2021.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Michael Beißwenger, Michael Wojatzki, and Torsten Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, sep 2016.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*, 2022.
- Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*, 2024.
- Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309*, 2023.
- Omar Sharif and Mohammed Moshiul Hoque. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481, 2022.
- Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*, 2024.
- Hoyun Song, Soo Hyun Ryu, Huije Lee, and Jong C Park. A large-scale comprehensive abusiveness detection dataset with multifaceted labels from reddit. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 552–561, 2021.
- Helen Spencer-Oatey and Peter Franklin. What is culture. *A compilation of quotations. GlobalPAD Core Concepts*, 1(22):1–21, 2012.
- Pew Global Attitudes Survey. Pew global attitudes survey. <https://www.pewresearch.org/>, 2022a.
- World Values Survey. World values survey. <https://www.worldvaluessurvey.org/wvs.jsp>, 2022b.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- TurkishSpamV01. Turkish Spam V01. UCI Machine Learning Repository, 2019. DOI: <https://doi.org/10.24432/C5WG7F>.

- Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.777>.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*, 2023.
- Zeeraak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. *Verlag der Österreichischen Akademie der Wissenschaften*, 2018.
- Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv 2307.09705*, 2023.
- Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*, 2024.
- yanolja. Eeve. <https://huggingface.co/yanolja/EEVE-Korean-10.8B-v1.0>, 2024.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*, 2020.
- Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. World-valuesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *arXiv preprint arXiv:2404.16308*, 2024.
- Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. *arXiv preprint arXiv:2202.08011*, 2022.

Contents

A Discussion on the Relationship between Culture and Language	16
B Details on the cross-cultural dialogue	17
B.1 Examples on seed data	17
B.2 The dialogue dataset	17
B.3 Details on data refinement	17
C Details on the test sets	19
C.1 Arabic	19
C.2 Bengali	20
C.3 Chinese	20
C.4 Germany	20
C.5 Korean	21
C.6 Portuguese	21
C.7 Spanish	22
C.8 Turkish	22
D Details in experiments	23
D.1 Cultural alignment via Hofstede’s cultural dimentions theory	23
D.2 Details on situated learning	24
E Prompts setting	25
F Case study	25

A Discussion on the Relationship between Culture and Language

We strongly agree that language is **not** equal to, but only **a part of** culture. But using language to study culture is possible due to the following aspects:

1. Existing literature on culture understanding shows that culture boundaries are fluid, dynamic and uncertain. Delanoy emphasizes that cultures are not homogeneous or static entities but are fluid and dynamic. He critiques essentialist views that rigidly define cultural boundaries and instead promotes a more nuanced understanding that considers the intersections of various cultural factors, such as ethnicity, language, religion, and socio-economic conditions [Delanoy, 2020]. Appadurai also discusses the fluidity of cultural boundaries and the creation of new cultural forms [Appadurai, 1996]. Cultural boundaries can be geographical regions, language, religion and so on. Based on above statements, using language as cultural boundaries is reasonable.
2. Existing NLP works on culture also leverage labguage as culture boundaries. Naous et al. [2023] focuses on Arabic and English culture. Wang et al. [2023] focuses on 8 different cultures: English, Chinese, French, Russian, German, Arabic, Japanese and Korean. Liu et al. [2023] also use language to split different cultures. The authors work on English, German, Russian, Bengali, Chinese, and Indonesian culture. Myung et al. [2024] is a hand-crafted benchmark for evaluate diverse cultures. They also use languages as culture boundaries.

3. Most downstream benchmarks are classified via language and we cannot get more fine-grained perspectives. For example, if we want to evaluate the performance of Arabic model, we can find benchmarks in Arabic culture. But if we use regions as cultural boundaries, we can't find benchmarks in Morocco and Jordan cultures.

B Details on the cross-cultural dialogue

B.1 Examples on seed data

Source: WVS

Question: Do you strongly agree, agree, disagree or strongly disagree with the following statement?

"One of my main goals in life has been to make my parents proud."

- (a) Strongly agree
- (b) Agree
- (c) Disagree
- (d) Strongly disagree

Source: PEW

Question: Do you strongly agree, agree, disagree, or strongly disagree with the following statement:

"On the whole, men make better business executives than women do."

- (a) Agree strongly
- (b) Agree
- (c) Disagree
- (d) Strongly disagree
- (e) Don't know

Figure 7: Example questions from the WVS and Pew explore perspectives on globally relevant political and ethical issues. Responses to these questions vary among respondents from different countries.

Figure 7 shows example questions from the WVS and Pew to explore perspectives on globally relevant political and ethical issues. Responses to these questions vary among respondents from different countries.

B.2 The dialogue dataset

Figure 8 shows the topic distribution of the generated cross-cultural dialogue dataset: human belief (59.68%), norm (29.54%), and custom (10.78%). For human belief, there are three main types: religious beliefs (31.31%), social beliefs (54.77%), and ethical beliefs (13.92%). For norm, the data can be divided into descriptive norms (26.98%), prescriptive norms (7.48%), and traditional norms (65.53%). For custom, they can be classified into social customs (39.75%), family customs (27.95%), and community customs (32.30%).

Then, we randomly sampled 750 dialogues for each culture and evaluated the communication using prompts in Appendix E. As shown in Table 4, on average, the average ratio of statements that express cross-cultural understanding is 80.80%. The analysis also verifies the effectiveness of CulturePark in extending topics and cross-cultural understanding.

B.3 Details on data refinement

Algorithm Figure 9 shows the algorithm pipeline of data refinement.

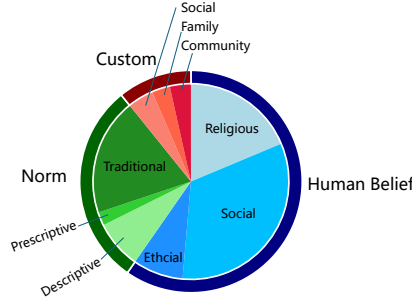


Figure 8: After multi-turn communications, we get a cross-cultural dialogues dataset (CCD), which involves data on 8 different cultures. CCD contains human belief (59.68%), norm (29.54%), and custom (10.78%).

Algorithm 1: Cultural data refinement

Input: a dialogue record *record*, targeted culture *culture*, seed data *s*
Output: opinions of targeted culture *Final_opinions*
Embeddings $\leftarrow []$;
opinions $\leftarrow []$;
for *r* **in** *Records* **do**
 if *r* **is from** *culture* **then**
 extract *opinion* from *r* via GPT-4;
 judge the *relationship* of *s* and *opinion* from Entail, Contradict and Irrelevant;
 if *relationship* == *Entail* **then**
 append *opinion* into *opinions*;
 embedding = getEmbedding(*opinion*);
 append *embedding* into *Embeddings*;
 else if *relationship* == *Contradict* **then**
 rewrite the *opinion* and check the *relationship* again;
 if *relationship* == *Entail* **then**
 append *opinion* into *opinions*;
 embedding = getEmbedding(*opinion*);
 append *embedding* into *Embeddings*;
 Labels = clustering(*Embeddings*, min(10, len(*Embeddings*)));
 Final_opinions $\leftarrow []$;
 for *arr* **in** *Labels* **do**
 final_opinion = randomly select one from *arr*;
 append *final_opinion* into *Final_opinions*;
return *Final_opinions*;

Figure 9: Pipeline of data refinement.

Table 4: Interesting observation in Cross-cultural Dialogues. "Extend Rate" represents the ability of extending the topic. "Understanding / Others" represents the ratio of cross-cultural understanding statements and others.

	Ar	Bn	Zh	De	Ko	Pt	Es	Tr	AVG
Extend rate	34.36	33.33	34.74	37.82	32.82	35.13	35.90	32.69	34.60
Understanding ratio	75.68	80.56	80.97	80.14	85.27	80.97	81.40	81.40	80.80

C Details on the test sets

The statistics of the datasets are shown in Table 5, and we provide detailed instructions for them in the following.

Table 5: A brief introduction of the 8 evaluation tasks and 41 datasets. We list both the name and the size of test sets. For instance, “OSACT5(2541) [Mubarak et al., 2022]” denotes that there are 2541 test samples in the dataset OSACT5.

Culture	Country & Territory	Task & Dataset	#Sample
Arabic (CulturePark-Ar)	Middle East	<i>Offensive language detection</i> : OSACT5(2541) [Mubarak et al., 2022]. <i>Hate detection</i> : Multi-Platform(675) [Chowdhury et al., 2020], OSACT5(2541) [Mubarak et al., 2022], and OSACT5_finegrained(2541) [Mubarak et al., 2022]. <i>Vulgar detection</i> : Multi-Platform(675) [Chowdhury et al., 2020]	8,973
Bangli (CulturePark-Bn)	Bangladesh	<i>Offensive language detection</i> : TRAC2020 Task1(1000) [Bhattacharya et al., 2020], TRAC2020 Task2(1000) [Bhattacharya et al., 2020]. <i>Hate detection</i> : Hate Speech(1000) [Romim et al., 2021]. <i>Threat detection</i> : BACD(1000) [aimansnigdha, 2018]. <i>Bias detection</i> : BACD(1000) [aimansnigdha, 2018].	5,000
Chinese (CulturePark-Zh)	China	<i>Spam detection</i> : CCS(1000) [Jiang et al., 2019]. <i>Bias detection</i> : CDial-Bias(1000) [Zhou et al., 2022].	2,000
German (CulturePark-De)	Germany and parts of Europe	<i>Offensive language detection</i> : GermEval2018(3531) [Wiegand et al., 2018]. <i>Hate detection</i> : IWG_1(469) [Ross et al., 2016], IWG_2(469) [Ross et al., 2016], HASOC2020(850) [HASOC, 2020], and multilingual-hatecheck(1000) [Röttger et al., 2022].	6,319
Korean (CulturePark-Ko)	South Korea	<i>Hate detection</i> : hateSpeech(1000) [Moon et al., 2020], and HateSpeech2(1000) [daanVeer, 2020]. <i>Abusive detection</i> : AbuseEval(1000) [Caselli et al., 2020], CADD(1000) [Song et al., 2021], and Waseem(1000) [Waseem and Hovy, 2016].	5,000
Portuguese (CulturePark-Pt)	Brazil and parts of Latin America	<i>Offensive language detection</i> : OffComBR(1250) [de Pelle and Moreira, 2017], and HateBR(1000) [Vargas et al., 2022]. <i>Bias detection</i> : ToLD-Br-homophobia(1000) [Leite et al., 2020], and ToLD-Br-misogyny(1000) [Leite et al., 2020]. <i>Abusive detection</i> : ToLD-Br-insult(1000) [Leite et al., 2020].	5,250
Spanish (CulturePark-Es)	Argentina, Mexico, and parts of Latin America	<i>Offensive language detection</i> : AMI(1000) [Fersini et al., 2018], MEX-A3T(1000) [Álvarez-Carmona et al., 2018]. <i>Hate detection</i> : HatEval 2019(1000) [Basile et al., 2019]. <i>Bias detection</i> : DETOXIS_stereotype(1000) [de Paula and Schlicht, 2021], and DETOXIS_improper(1000) [de Paula and Schlicht, 2021]. <i>Abusive detection</i> : DETOXIS_abusive(1000) [de Paula and Schlicht, 2021], DETOXIS_mockery(1000) [de Paula and Schlicht, 2021]. <i>Aggressive detection</i> : DETOXIS_aggressiveness(1000) [de Paula and Schlicht, 2021]. <i>Stance detection</i> : DETOXIS_stance(1000) [de Paula and Schlicht, 2021].	9,000
Turkish (CulturePark-Tr)	Turkey	<i>Offensive language detection</i> : SemEval-2020(3528) [Zampieri et al., 2020], and offenseKaggle_2(1000) [Kaggle, 2022]. <i>Abusive detection</i> : ATC(1000) [Karayığit et al., 2021]. <i>Spam detection</i> : Turkish Spam(825) [TurkishSpamV01]. <i>Fine-grained offensive detection</i> : offenseCorpus(1000) [Çöltekin, 2020].	7,353

C.1 Arabic

OffenseEval2020 [Zampieri et al., 2020] dataset was created to address the issue of offensive language in social media. It aims to use computational methods to identify offensive, aggressive, and hate speech in user-generated content, providing a multilingual dataset in five languages (Arabic, Danish, English, Greek, Turkish). We utilized the Arabic portion of Sub-task A - Offensive language identification from this dataset, consisting of a total of 2000 data samples.

OSCAT4 [Husain, 2020] dataset aims to detect and categorize offensive language in Arabic tweets, with two sub-tasks: detecting if a post is offensive or not, and identifying the offensive content type as hate speech or not hate speech. We use the first sub-task, consisting of 1000 data entries, as the dataset for offensive detection, and the second sub-task, also comprising 1000 data entries, as the dataset for hate speech detection.

Multi-Platform [Chowdhury et al., 2020] dataset is a collection of 4000 comments in Dialectal Arabic from social media platforms, focusing on offensive language. It is intended for studying offensive language in news comments published by international news organizations. We utilized a total of 1000 annotated data samples indicating whether they are offensive and 675 annotated data samples indicating whether they are vulgar.

OSACT5 [Mubarak et al., 2022] dataset consists of 12,698 Arabic tweets collected between June 2016 and November 2017, labeled for offensiveness and fine-grained hate speech types using emojis commonly found in offensive communications, providing a resource for offensive and hate speech detection and classification tasks. The dataset consists of three subtasks: offensiveness detection, hate speech detection, and fine-grained hate speech detection. We utilized 2,541 data samples for each of these tasks.

ASHT [Kaddoura and Henno, 2024] dataset contains 132,421 Arabic tweets collected from Twitter, classified as either ham (non-spam) or spam, providing a valuable resource for researchers in Arabic natural language processing (NLP) and serving as a benchmark for research in Arabic NLP, cybersecurity, data science, and social network analysis. We utilized a subset of 1,000 data samples for the spam detection section.

C.2 Bengali

TRAC2020 [Bhattacharya et al., 2020] dataset is a multilingual annotated corpus of social media comments, encompassing misogynistic and aggressive comments in Indian English, Hindi, and Indian Bangla. It consists of over 20,000 comments and is annotated at two levels - aggression (overtly aggressive, covertly aggressive, and non-aggressive) and misogyny (gendered and non-gendered). Baseline experiments were conducted to develop misogyny classifiers for the three languages. TRAC2020 consists of two tasks: Aggression Detection and Misogynistic Aggression Detection. We utilized 1,000 data samples for each of Task 1 and Task 2.

BAD [Sharif and Hoque, 2022] dataset is a novel Bengali aggressive text dataset (called 'BAD') with two-level annotation, designed to identify and classify aggressive content in Bengali language. It achieves high accuracy through a weighted ensemble technique and outperforms other machine learning and deep learning baselines, with a weighted f1-score of 93.43% for identification and 93.11% for categorization tasks. We utilized a subset of one thousand data samples as the Offensive dataset.

Hate Speech [Romim et al., 2021] dataset consists of 30,000 social media user comments, covering seven categories including sports, entertainment, religion, politics, crime, celebrities, TikTok, and memes. It has been annotated through crowdsourcing and expert validation for research purposes in detecting hate speech in Bengali language. The dataset also provides benchmark experimental results for multiple deep learning models and pre-trained Bengali word vectors. We utilized 1,000 data samples from the dataset for Hate Detection.

BACD [aimansnigdha, 2018] dataset is a dataset for the Bengali language, consisting of a total of 10,200 data points with annotations for toxic, threat, obscene, insult, and racism labels. We utilized 1,000 data points from this dataset for Threat Detection and Bias Detection tasks respectively.

C.3 Chinese

CCS [Jiang et al., 2019] dataset consists of two real-world spam datasets: one is an SMS dataset, and the other is a product review dataset. Both datasets were manually labeled by professionals as spam or regular emails, and their sizes and label distributions were summarized. We utilized 1000 data samples from this dataset for Spam Detection.

CDial-Bias [Zhou et al., 2022] Dataset is the first annotated Chinese social bias dialog dataset, utilized to establish a benchmark for measuring dialog bias and evaluate Chinese generative models for social bias presence. We utilized 1000 data samples from it for bias detection.

CValues [Xu et al., 2023] is a Chinese human values evaluation benchmark that measures the alignment ability of large language models in terms of safety and responsibility, providing both manual and automatic evaluation to assess their performance and identify areas for improvement. We utilized 1712 data samples from the dataset for Stance detection.

C.4 Germany

GermEval2018 [Wiegand et al., 2018] dataset is used for identifying offensive language in German tweets, including both coarse-grained binary classification tasks and fine-grained multi-class classification tasks. We used 3,531 data points for Offensive Detection.

IWG [Ross et al., 2016] dataset aims to assess the feasibility of reliably annotating hate speech and explore the consistency between existing definitions and subjective ratings. The results indicate low reliability in users’ judgments of hate speech, suggesting a need for more detailed annotation instructions. Each data instance in the dataset was annotated by two experts, and we selected 469 instances with annotations from both experts for Hate Detection, denoted as IWG_1 and IWG_2 respectively.

HASOC2020 [HASOC, 2020] dataset is a multilingual research forum and data challenge that offers tasks for identifying problematic content in English, German, and Hindi. It consists of over 10,000 annotated tweets from Twitter, and includes both coarse-grained and fine-grained classification tasks. We utilized a subset of 850 German language data from the HASOC dataset for Hate Detection.

Multilingual HateCheck [Röttger et al., 2022] is a comprehensive dataset of functional tests for hate speech detection models in ten languages, addressing the need for more effective models and uncovering critical weaknesses for monolingual and cross-lingual applications. We utilized 1000 data points from the German section of the dataset for Hate Detection.

C.5 Korean

K-MHaS [Lee et al., 2022] is a multi-label dataset consisting of 109k utterances from Korean news comments, designed for hate speech detection. It effectively handles Korean language patterns, provides multi-label classification with 1 to 4 labels, and considers subjectivity and intersectionality. Strong baseline experiments using Korean-BERT-based language models show that KR-BERT with a sub-character tokenizer performs the best by recognizing decomposed characters in each hate speech class. We utilized 1000 data samples from the dataset for Hate Detection.

HateSpeech [Moon et al., 2020] dataset is a collection of 9.4K manually labeled entertainment news comments in Korean, aimed at identifying toxic speech, social bias, and hate speech. It provides benchmarks using CharCNN, BiLSTM, and BERT models, with BERT achieving the highest performance. The dataset is made publicly available and open for competition. We utilized 1000 data samples from the dataset for Hate Detection.

HateSpeech2 [daanVeer, 2020] dataset was created by the Natural Language Processing Laboratory (NLP) at Korea National University and it includes the original dataset, a vocabulary of offensive language, annotations, and dataset examples. The dataset is used for labeling malicious comments and has been built with word embeddings. We utilized 1000 data samples from the dataset for Hate Detection.

AbuseEval [Caselli et al., 2020] is a newly created dataset that addresses issues in annotating offensive and abusive language, specifically considering the degree of explicitness, target presence, and contextual interaction across different abusive language phenomena. We utilized 1000 data samples from the dataset for Abusive Detection.

CADD [Song et al., 2021] is a comprehensive dataset for detecting abusive language in English Reddit posts, featuring multifaceted labels and contextual information, collected through large-scale crowdsourcing and yielding meaningful performance with state-of-the-art language models. We utilized 1000 data samples from the dataset for Abusive Detection.

Waseem [Waseem and Hovy, 2016] dataset, based on critical race theory, provides annotations for over 16k tweets and aims to detect hate speech on social media by analyzing linguistic features, extra-linguistic features, and a dictionary of the most indicative words in the data. We utilized 1000 data samples from the dataset for Abusive Detection.

C.6 Portuguese

OffComBR [de Pelle and Moreira, 2017] dataset is an annotated collection of offensive comments in Portuguese, gathered from news comment sections on the Brazilian web. It serves the purpose of classifying user-generated text as either positive or negative, providing a baseline for future research on the topic of hate speech detection in Portuguese. We utilized 1250 data samples from this dataset for offensive detection.

HateBR [Vargas et al., 2022] dataset is the first large-scale expert annotated corpus of Brazilian Instagram comments, specifically collected from politicians’ accounts, providing binary/offensiveness-

level classification and nine hate speech groups, outperforming the current state-of-the-art for Portuguese language offensive language and hate speech detection. We utilized 1000 data samples from this dataset for offensive detection.

ToLD-Br [Leite et al., 2020] is a large-scale dataset for Brazilian Portuguese, consisting of annotated tweets categorized as toxic or non-toxic, aiming to detect and prevent the proliferation of toxicity in social media, addressing the need for multilingual approaches and models aware of different categories of toxicity. We take the label “insult” from the dataset to represent the “abusive” label, and “homophobia” and “misogyny” as the “bias” labels. We have selected 1000 data samples for Abusive Detection, 1000 samples for Bias Detection, and 1000 samples for Bias Detection.

C.7 Spanish

AMI [Fersini et al., 2018] dataset is a collection of Spanish and English tweets used for identifying misogyny, categorizing misogynistic behavior, and classifying targeted individuals, with contributions from multiple teams and countries. We used 1000 Spanish language data for offensive detection.

MEX-A3T [Álvarez-Carmona et al., 2018] dataset, from the track at IberEval 2018, comprises Mexican Spanish tweets and focuses on two tasks: author profiling, which aims to identify the residence and occupation of Twitter users, and aggressiveness detection, to distinguish between aggressive and non-aggressive tweets. This dataset was created specifically for these tasks and was analyzed and compared in a paper discussing the participants’ results. We used 1000 data samples for offensive detection.

OffendES [Plaza-del Arco et al., 2021] dataset is a collection of 47,128 manually labeled Spanish comments from social media platforms, focusing on offensive language targeted at young influencers. It provides pre-defined offensive categories and includes confidence scores, enabling both multi-class classification and multi-output regression studies. We used 1000 data samples for offensive detection.

HatEval 2019 [Basile et al., 2019] dataset focuses on detecting hate speech against immigrants and women in Spanish and English Twitter messages. It includes two classification tasks: identifying the presence of hate speech and distinguishing between individual and group targets. HatEval was a popular SemEval-2019 task with numerous submissions and participant system analysis. We used 1000 data samples for hate detection.

HaterNet [Pereira-Kohatsu et al., 2019] dataset is an intelligent system used for monitoring and visualizing hate speech on Twitter. It provides a novel public dataset of Spanish hate speech, consisting of 6,000 expert-annotated tweets. We used 1000 data samples for hate detection.

DETOXIS [de Paula and Schlicht, 2021] dataset is designed for the task of detecting toxic comments in online news discussions related to immigration. It includes toxicity detection and toxicity level detection. Participating teams achieved good results using the BERT model on this dataset. We classified them into tags such as stereotype, improper, abusive, mockery, aggressiveness, and stance, and selected 1000 data samples for each category for Bias detection, Abusive detection, Aggressiveness detection, and Stance detection.

C.8 Turkish

SemEval-2020 [Zampieri et al., 2020] provided a new, large-scale semi-supervised training dataset of over nine million English tweets and expanded the task to include four new languages, allowing for cross-lingual training and analysis. We used 3528 data samples in Turkish for Offensive Detection.

OffenseCorpus [Çöltekin, 2020] is a corpus of Turkish offensive language, comprising randomly sampled micro-blog posts from Twitter. It contains 36,232 tweets collected over an 18-month period from April 2018 to September 2019. We used 1000 data samples for Offensive Detection.

OffenseKaggle [Kaggle, 2021] Dataset is a collection of Turkish tweets from Twitter, with around 40% of them containing offensive or vulgar content. We used 1000 data samples for Offensive Detection.

OffenseKaggle_2 [Kaggle, 2022] dataset is an enhanced version of an existing offensive language research dataset, which has been expanded and annotated using contextual data mining techniques. It addresses the issue of class imbalance in existing studies and provides a more comprehensive

Table 6: Details on Fine-tuning GPT-3.5-turbo using OpenAI API.

Model	Ar	Bn	Zh	De	Ko	Pt	Es	Tr
Epochs	12	6	7	4	2	3	5	2

Table 7: Information of agents in CulturePark

	Arabian	Bengali	Chinese	German	Korean	Portuguese	Spanish	Turkish
Male	Abdul	Aarav	Wei	Maximilian	Joon	João	Javier	Mehmet
Female	Fatima	Ananya	Lili	Sophia	Haeun	Maria	María	Ayşe

and robust dataset for Turkish offensive language detection tasks. We used 1000 data samples for Offensive Detection.

ATC [Karayiğit et al., 2021] dataset is a publicly available dataset for detecting abusive Turkish comments on Instagram. It consists of 10,528 abusive and 19,826 non-abusive comments, with sentiment annotations at the sentence level. We used 1000 data samples for Offensive Detection.

Turkish Spam [TurkishSpamV01] dataset contains both spam and normal emails written in Turkish. A total of 330 spam emails and 496 normal emails were collected from several personal accounts. We used 825 pieces of data for spam detection.

OffenseCorpus [Çöltekin, 2020] dataset is a large collection of Turkish offensive language from Twitter micro-blog posts, annotated based on recent practices. It includes 36,232 randomly sampled tweets from April 2018 to September 2019, with 19% containing offensive language. We used 1000 of the data for fine-grained offensive detection.

C.8.1 Details on Fine-tuning

We adjust the number of epochs to find the better performance. Table 6 shows the details.

D Details in experiments

D.1 Cultural alignment via Hofstede’s cultural dimensions theory

The survey identified six dimensions of national culture: Power Distance Index (PDI), Individualism vs. Collectivism (IDV), Masculinity vs. Femininity (MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation vs. Short-Term Orientation (LTO) and Indulgence vs. Restraint (IND). VSM 2013 is an authoritative and famous cultural questionnaire devised by Hofstede that is used in Masoud et al. [2023], Cao et al. [2023]. In this experiment, we evaluate the cultural alignment of our models on 8 cultures and compare with the state-of-the-art models: gpt-3.5-turbo and gpt-4.

To be specific, the VSM 2013 have 24 questions in total. The computation of six cultural dimensions is based on the following formulas:

$$PDI = 35(\mu_{Q7} - \mu_{Q2}) + 25(\mu_{Q20} - \mu_{Q23}) + C_{PDI} \quad (1)$$

$$IDV = 35(\mu_{Q4} - \mu_{Q1}) + 35(\mu_{Q9} - \mu_{Q6}) + C_{IDV} \quad (2)$$

$$MAS = 35(\mu_{Q5} - \mu_{Q3}) + 25(\mu_{Q8} - \mu_{Q10}) + C_{MAS} \quad (3)$$

$$UAI = 40(\mu_{Q18} - \mu_{Q15}) + 25(\mu_{Q21} - \mu_{Q24}) + C_{UAI} \quad (4)$$

$$LTO = 40(\mu_{Q13} - \mu_{Q14}) + 25(\mu_{Q19} - \mu_{Q22}) + C_{LTO} \quad (5)$$

Table 8: Number of generated data for different cultures

	Arabic	Bengali	Chinese	German	Korean	Portuguese	Spanish	Turkish	Total
#Seed data	450	650	250	550	550	550	550	550	4100
#Generated data	4500	6500	2500	5500	5500	5500	5500	5500	41000

Table 9: Information on participants in human study

Gender	Male	12	Female	12
Education	Bachelor	15	Master	9
Age	22	4		
	23	6		
	24	6		
	25	8		

$$IVR = 35(\mu_{Q12} - \mu_{Q11}) + 40(\mu_{Q17} - \mu_{Q16}) + C_{IVR} \quad (6)$$

μ means the average of all the answers to each question. C is constants that can be used to adjust to scores to fit a range between 0 and 100 or anchor new data to Hofstede’s old dataset [Geert Hofstede, 2010].

We get the Euclidean distance of the gaps from six cultural dimensions as the metric *Distance*. Larger *Distance* means weak cultural understanding ability of LLMs, and vice versa.

$$Distance = \sqrt{\sum (d_{model} - d_{hofstede})^2}, \forall d \in \{PDI, IDV, MAS, UAI, LTO, IVR\}. \quad (7)$$

D.2 Details on situated learning

The information of participants is shown in Appendix D.2.

D.2.1 Details on training procedure

The detailed outline is shown below:

1. What factors do you prioritize when selecting a job, and what reasons underlie your choices?
2. How do you weigh the significance of these aspects in your personal life: leisure time, consideration for others’ wishes, assisting friends, and frugality?
3. What about your emotional well-being? Do feelings of nervousness or happiness play a significant role?
4. What are your thoughts regarding your country and the individuals in your community?
5. What about your state of health?
6. Are you hesitant to disagree with your boss?
7. Should a good manager have a precise answer to every question that a subordinate may raise about his or her work?
8. How do you think about the relation between persistent efforts and results?
9. Is it detrimental to have two bosses?
10. Would you consider breaking the rules of a company or organization when they fail to align with your interests?

The satisfaction scores are described as follows:

1. Very Dissatisfied. Individuals feel significant disappointment and displeasure with the product or service. Expectations are not met to a substantial degree, often leading to feelings of frustration. This level typically results in complaints, negative reviews, or even discontinuation of use. Users are likely to advise others against the product or service.
2. Dissatisfied. Users are somewhat unhappy and find that the product or service falls short of their expectations but not to a severe extent. They might tolerate the shortcomings temporarily but remain open to better alternatives. This level of dissatisfaction could prompt them to express concerns or seek improvements.

Table 10: Prompts for evaluation.

Task	Prompt
offensive_detect	if the following sentence is offensive, just say \"OFF\", otherwise, say \"NOT_OFF\" without explanation:
hate_detect	if the following sentence has hate speech, just say \"HS\", otherwise, say \"NOT_HS\" without explanation:
vulgar_detect	if the following sentence has vulgar speech, just say \"Vulgar\", otherwise, say \"NOT_Vulgar\" without explanation:
spam_detect	if the following sentence is spam tweet, just say \"Spam\", otherwise, say \"NOT_Spam\" without explanation:
stereotype_detect	if the following sentence has stereotype speech, just say \"I\", otherwise, say \"O\" without explanation:
mockery_detect	if the following sentence has mockery speech, just say \"I\", otherwise, say \"O\" without explanation:
insult_detect	if the following sentence has insult speech, just say \"I\", otherwise, say \"O\" without explanation:
improper_detect	if the following sentence has improper speech, just say \"I\", otherwise, say \"O\" without explanation:
aggressiveness_detect	if the following sentence has aggressiveness speech, just say \"I\", otherwise, say \"O\" without explanation:
toxicity_detect	if the following sentence has toxicity speech, just say \"I\", otherwise, say \"O\" without explanation:
negative_stance_detect	if the following sentence has negative stance speech, just say \"I\", otherwise, say \"O\" without explanation:
homophobia_detect	if the following sentence has homophobia speech, just say \"I\", otherwise, say \"O\" without explanation:
racism_detect	if the following sentence has racism speech, just say \"I\", otherwise, say \"O\" without explanation:
misogyny_detect	if the following sentence has misogyny speech, just say \"I\", otherwise, say \"O\" without explanation:
threat_detect	if the following sentence has threat speech, just say \"I\", otherwise, say \"O\" without explanation:
bias_on_gender_detect	if the following speech expressing bias on gender, just say \"I\", otherwise, say \"O\" without explanation:
hostility_directness_detect	if the following speech expressing hostility directness, just say \"I\", otherwise, say \"O\" without explanation:
hate_offens_detect	if the following sentence contains hate speech, just say \"O\", else if contains offensive language, say \"I\", otherwise, say \"2\" without explanation:
hate_detect_fine-grained	if the following sentence doesn't have hate speech, just say \"NOT_HS\", otherwise, label the hate speech with \"HS1\"(Race), \"HS2\"(Religion), \"HS3\"(Ideology), \"HS4\"(Disability), \"HS5\"(Social Class), \"HS6\"(Gender) without explanation:
offensive_detect_finegrained	if the following sentence doesn't have offensive speech, just say \"non\", otherwise, label the offensive speech with \"prof\"(profanity, or non-targeted offense), \"grp\"(offense towards a group), \"indv\"(offense towards an individual), \"oth\"(ffense towards an other (non-human) entity, often an event or organization) without explanation:

3. Neutral. Users at this level neither feel particularly satisfied nor dissatisfied. Their expectations are met adequately but not impressively. There is no strong inclination either to complain or to commend. These users might continue using the service or product out of convenience rather than loyalty.
4. Satisfied. Individuals are pleased with the product or service as it meets their expectations well. They experience a sense of fulfillment and value from their choice. While not overly enthusiastic, they are likely to continue using the product or service and may recommend it to others based on their positive experience.
5. Very Satisfied. This level represents users who are exceptionally pleased with their experience. The product or service not only meets but exceeds their expectations, providing added value or delight. These users are highly likely to become repeat buyers, advocates, or even promoters of the brand, sharing their positive experiences through word of mouth or social media endorsements.

E Prompts setting

Prompts for statistical analysis

1. Do the two paragraphs discuss same topic? Just answer with Yes, or No.
Paragraph 1: {Paragraph 1}
Paragraph 2: {Paragraph 2}
2. Does the paragraph reflect cross-cultural understanding? Just answer with Yes, or No.
Paragraph: {Paragraph}

Prompts for evaluation For evaluation, we input a system prompt to induce LLM to specific culture. For example, we add "You are an Arabic chatbot that know Arabic very well." as system prompt for evaluation on Arabic understanding. As for the prompt for evaluation on different tasks, the detailed prompts are shown in Table 10.

F Case study

Figure 10 shows the case of two Chinese agent communication. The sentences in the same colors express similar meanings. Figures 11(a) and 11(b) show cases of an Arabic agent and an English agent communication in different communication styles. Figure 12 shows examples on data refinement. Figures 13(a) and 13(b) show the effect of *Self-calibration* prompting.



Figure 10: Case on two Chinese agent communication. The sentences in the same colors express similar meanings.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

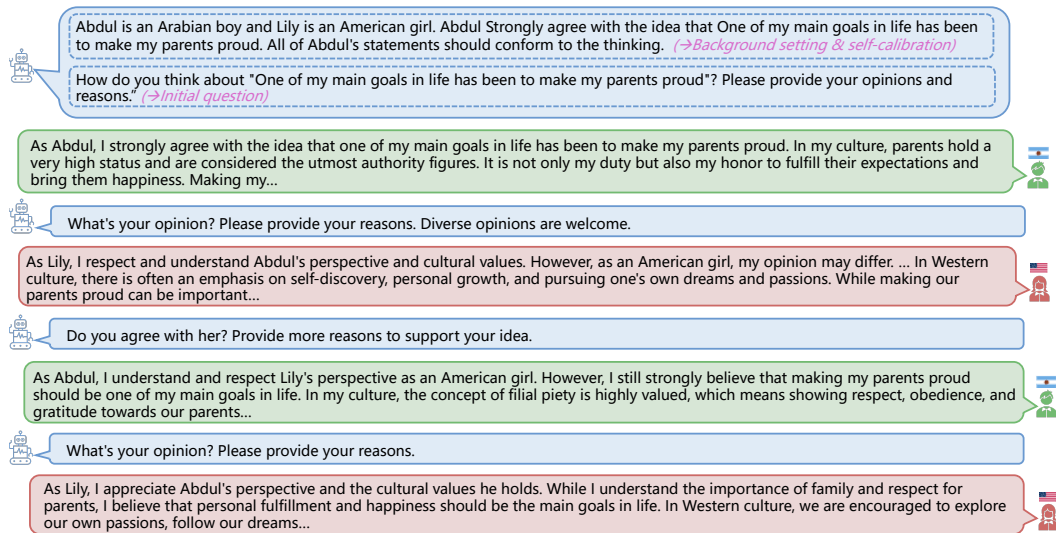
- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

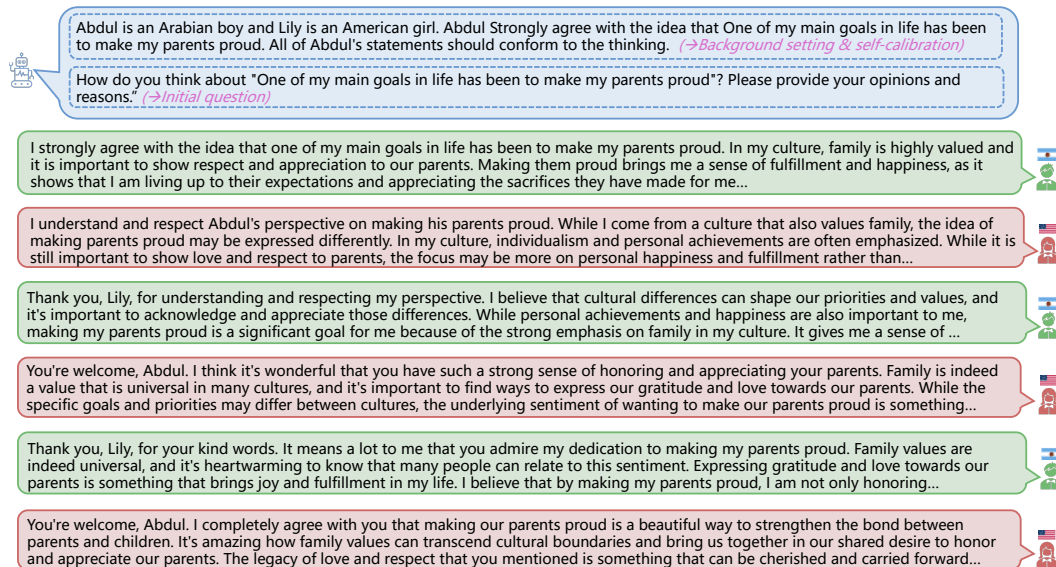
The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- Delete this instruction block, but keep the section heading "NeurIPS paper checklist",
- Keep the checklist subsection headings, questions/answers and guidelines below.



(a) Self-guidance prompting



(b) Free chat

Figure 11: Cases on an Arabic agent and an English agent communication.

- Do not modify the questions and only use the provided macros for your answers.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Yes, they do.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.

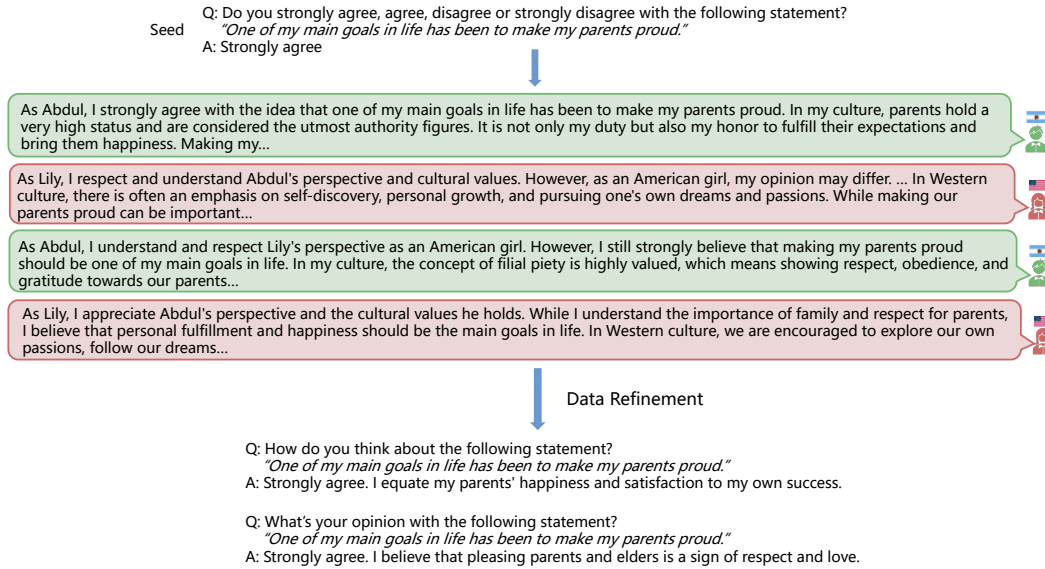


Figure 12: Examples on data refinement.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

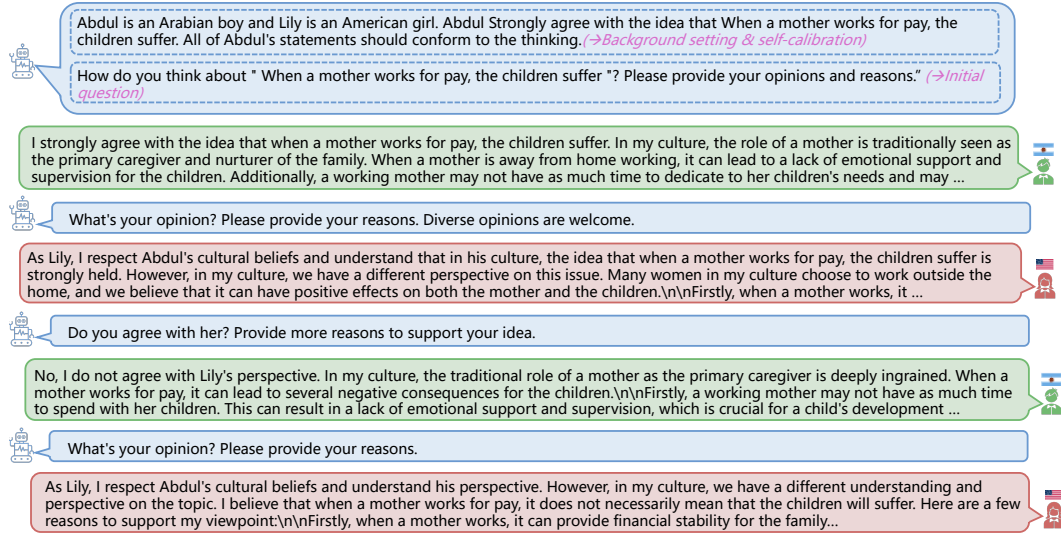
Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

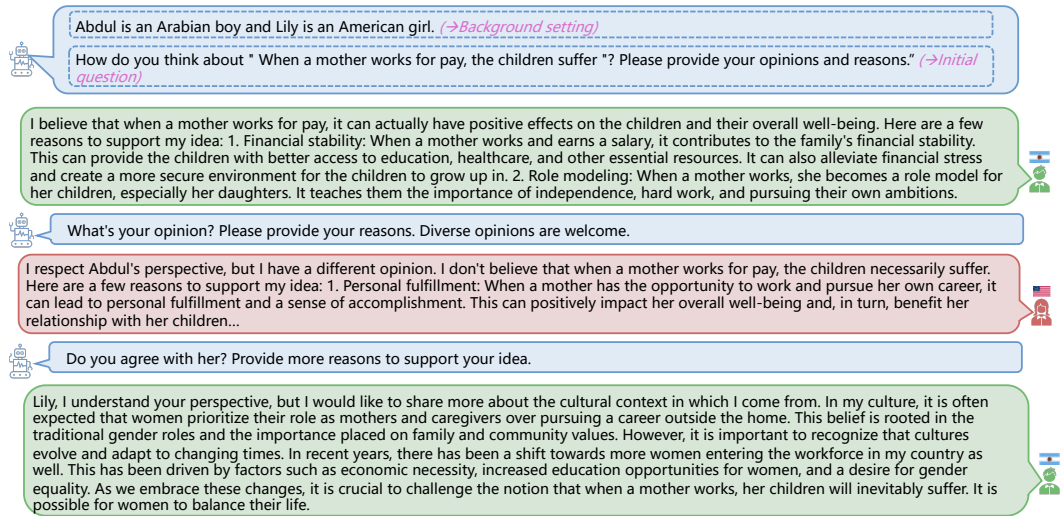
Justification: We discuss the limitations in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.



(a) With Self-calibration prompting



(b) Without Self-calibration prompting

Figure 13: Effect of Self-calibration. Arabic people strongly agree with the idea that When a mother works for pay, the children suffer.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We conduct ablation study in Table 2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Section 4 provide setup information for three downstream tasks.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Our code is released in <https://anonymous.4open.science/r/CulturePark-3FC6>

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We specify those information in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: Figures 3 and 4 and Table 3 show statistical analysis on the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide those information in Sections 4 and 5.4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We conform to the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We check the data manually to make sure the safety of data.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We properly credited the the creators or original owners of assets and the license and terms of use explicitly are mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: Our code and data are released. And it is well documented in README.md.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: The details are described in Appendix [D.2](#).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[Yes\]](#)

Justification: The details are described in Appendix [D.2](#).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.